WATER ANALYSIS WITH THE HELP OF TENSOR CANONICAL DECOMPOSITIONS

J.-P. Royer, P. Comon*

Université de Nice Sophia-Antipolis,
Lab. I3S, UMR6070
2000, route des Lucioles, BP.121
06903 Sophia Antipolis Cedex, France
pcomon@unice.fr

S. Mounier, R. Redon, H. Zhao

Université du Sud Toulon Var,
Lab. PROTEE, EA 3819
BP 20132
83957 La Garde Cédex, France
{mounier,roland.redon}@univ-tln.fr

N. Thirion-Moreau

Université du Sud Toulon Var,
Lab. LSEET, UMR CNRS 6017
Av. G. Pompidou, BP.56, 83162
La Valette du Var, Cédex, France
thirion@univ-tln.fr

C. Potot, G. Féraud

Université de Nice Sophia Antipolis,
Lab. Radiochimie, Sciences Analytiques
et Environnement, ICN, FR CNRS 3037,
Parc Valrose, 06108 Nice cedx 02
Cecile.POTOT@unice.fr

1. INTRODUCTION

This study has been started two years ago by the laboratories of Radiochimie, Sciences Analytiques et Environnement (LRSAE) of the University of Nice Sophia-Antipolis (UNS) and PROcessus de Transferts et d'Echanges dans l'Environnement (PROTEE) of the University of Sud Toulon-Var (USTV). It has numerous goals, among which the study of the exchanges between different kinds of rocks and soils (limestone, conglomerates, alluvial and phreatic layers), water and underground water to determine the impact of erosion phenomena, the nature of exchanges between the Var and its affluents (Vésubie and Esteron), the quality of water, the detection of potential polluting agents (marking elements and more precisely heavy metals like Pb, As and Co for example).

Since january 2009, water samples have been collected on a weekly basis in five locations, named: Var river (1), Auda (2), Maccario (3), Puget (4) and La Tour (5). Different measures are then performed: dissolved organic carbon, dissolved oxygen, pH, temperature, concentration of ions and heavy metals.

*Funded by a PhD support delivered by the University of Nice, in the frame of the PRES euro-méditerrannéen.

2. PROBLEM STATEMENT

2.1. Experimental configuration and its aims

The raw data are collected in five measurement locations. Their geographic position is depicted in Fig.1. It is assumed that some locations interact with each other, whereas others do not. In such a context, we are interested in determining the contribution of each location and in better understanding the water exchanges that are involved. Organic components can also be identified thanks to methods such as Canonical Polyadic decompositions (CP) (sometimes known as Parafac), applied to 3D fluorescence spectra calculated from the collected samples. Thus, organic elements will be tracked along the river.

2.2. Mathematical model and assumptions

Considering the aforementioned experimental configuration, we have set up a mathematical model, whose aim is to model the water exchanges between the chosen locations. It leads to various partial relations between data $C^{(i)}(t)$, measured at location i and time t, *e.g.* concentrations. For example, regarding area numbered 4, we have exchanges with areas 1, 3 and 5, but not with area 2, so that we can assume the model below:

$$\begin{aligned}\alpha_{44}C^{(4)}(t) = &\alpha_{41}C^{(1)}(t-\tau_{41}) + \alpha_{43}C^{(3)}(t-\tau_{43}) \\ &+ \alpha_{45}C^{(5)}(t-\tau_{45}) + \chi^{(4)}(t) \qquad (1)\end{aligned}$$

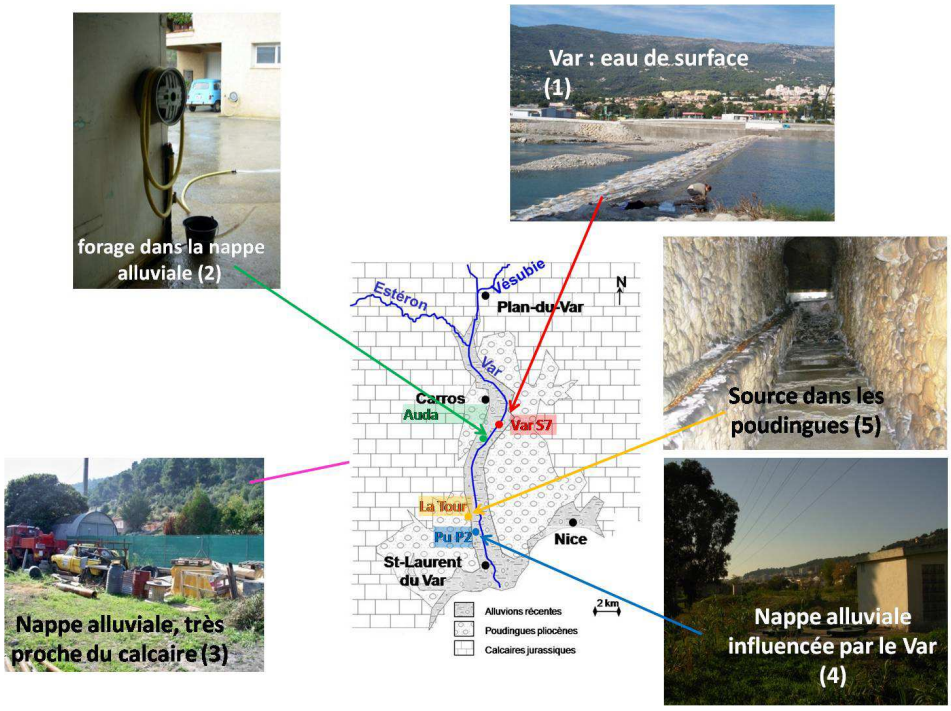

Fig. 1.

where α_{ij} stands for the flow from location j to i if $i \neq j$ and $\alpha_{ii} = \sum_{j \neq i} \alpha_{ij}$, τ_{ji} stands for the transport delay from site i to site j, and χ^i denotes an error term. We can do the same for area 3:

$$\alpha_{33}C^{(3)}(t) = \alpha_{31}C^{(1)}(t-\tau_{31}) + \alpha_{32}C^{(2)}(t-\tau_{32}) - \alpha_{23}C^{(3)}(t-\tau_{23}) + \chi^{(3)}(t) \quad (2)$$

and for area 2:

$$\alpha_{22}C^{(2)}(t) = \alpha_{21}C^{(1)}(t-\tau_{21}) + \alpha_{23}C^{(3)}(t-\tau_{23}) - \alpha_{32}C^{(2)}(t-\tau_{32}) + \alpha_{26}C^{(6)}(t) + \chi^{(2)}(t) \quad (3)$$

One can notice that flows can go in both directions, as it is observed for areas 2 and 3. In the equation above, $\alpha_{26}C^{(6)}(t)$ represents the contribution of the phreatic layer, which could have been be merged in the error term, since it cannot be measured.

As pointed out earlier, $C^{(i)}(t)$ represents some measurement performed at site i and time t. Assume for instance that it represents a fluorescence intensity (but it could be another type of measurement such as pH, etc). It can be decomposed as: $C(t) = \sum_{p=1}^{N_p} c_p(t)S_p$, where N_p denotes the (unknown) number of components in the mixing, $c_p(t)$ the concentration of the p^{th} component and S_p its fluorescence 3D spectrum. Actually, S_p is an intensity, which is measured as a function of emission and excitation wavelengths. So it is a function of two variables (as $C(t)$ in the case of fluorescence analysis); for the sake of simplicity, this dependence has not been made explicit in the notation. See the next section for more details.

Assumptions. Our subsequent developments are based on the assumptions below:

A1. Only conservative elements can be considered, in order to be able to estimate transport delays.

A2. Coefficients α_{ij} are constant for each component over the observation duration.

A3. Delays τ_{ij} are constant too, for each component.

Goals. One of our objectives is to estimate the transport delays τ_{ij}. To determine theses delays, one very basic idea is to search for maxima of inter-correlations between data from two linked locations. Additional information like the marking elements (*i.e.* metals) concentrations should help us to achieve this task.

The other goal is to determine the flows α_{ij}, which should provide a good estimation of the contribution of each location to the global system. We can already point out some of the difficulties that we have encountered. The estimation of the delays cannot be performed directly since measurements are not performed regularly (sparse sampling) and not synchronized (since not performed at the same time). Our first objective has been to resample the data on a regular grid as

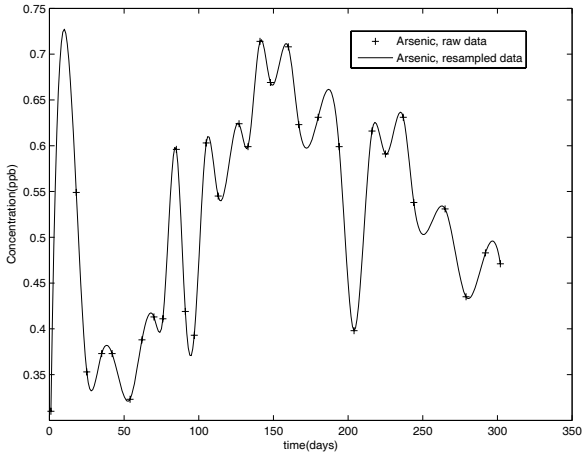

Fig. 2. Arsenic concentrations measured during one year (+) at site 4 (Puget), surperimposed on the resampled data (straight line).

illustrated by Fig. (2). Moreover, some technical problems can appear during the process of estimation of concentrations. This is why measurements are sometimes missing. This problem has to be taken into account too, especially when it concerns the first(s) or last(s) measurements of the considered time series.

Another difficulty to overcome is the fact that most measurement techniques do not provide us with an information on a single element, but on a mixture of elements. The goal is then to recover individual information from mixtures. This is addressed in the next section.

3. CANONICAL DECOMPOSITION

In order to fix the ideas, consider the case of a fluorescence analysis. If a solution is excited by an optical excitation, several effects may be produced: Rayleigh diffusion, Raman diffusion, and fluorescence. At low concentrations, the Beer-Lambert law can be linearized so that the fluorescence intensity rather accurately follows the model below [7]:

$$I(\lambda_f, \lambda_e, k) = I_o\, \gamma(\lambda_f)\, \epsilon(\lambda_e)\, c_k$$

where ϵ denotes absorbance spectrum (sometimes called excitation spectrum), λ_e the excitation wavelength, γ the fluorescence emission spectrum, λ_f is the fluorescence emission wavelength, and k denotes the sample number (e.g. which can vary concentration). Provided it can be separated from diffusion phenomena, the fluorescence phenomenon allows to determine the concentration of a diluted (fluorescent) chemical component, and possibly to recognize it thanks to its fluorescent spectrum.

A difficulty appears when the solution contains more than one fluorescent solute. In such a case, the overall fluorescence intensity is an unknown linear combination of component fluorescence intensities:

$$I(\lambda_f, \lambda_e, k) = I_o \sum_{\ell} \gamma_\ell(\lambda_f)\, \epsilon_\ell(\lambda_e)\, c_{k,\ell} \qquad (4)$$

It is then necessary to separate each component contribution.

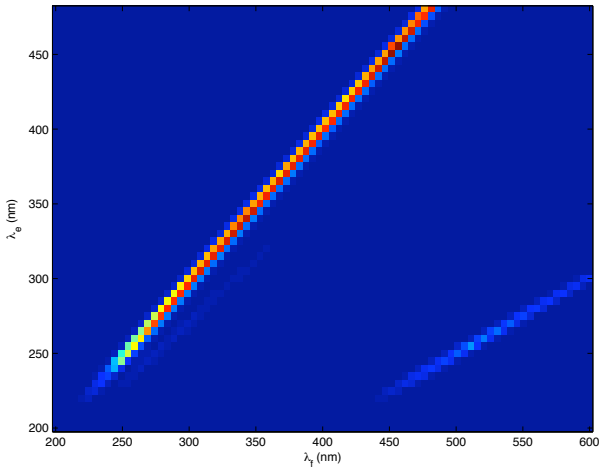

Fig. 3. 3D fluorescence spectrum of a water sample of Var river (Auda), before removal of Rayleigh and Raman effects; horizonal: λ_f, vertical: λ_e.

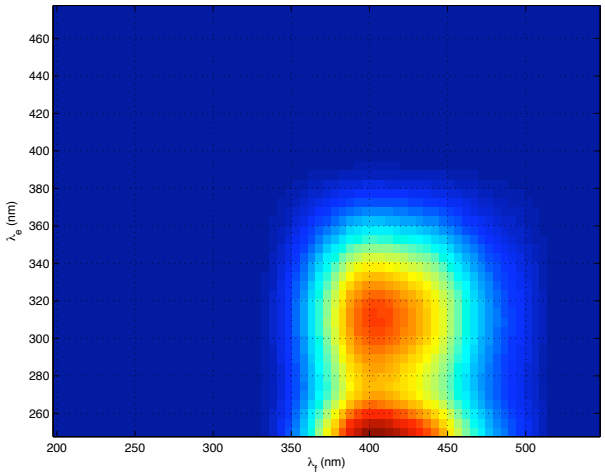

Fig. 4. After isolation of a 3D fluorescence spectrum, a component may be identified; here a PicM component, marine humic-like matters, Coble, 1996.

There exist a wide panel of separation techniques, allowing to identify linear mixtures of functions (or stochastic processes) and to extract them. Most of them rely on

statistical tools, or on sparsity; see for instance the survey provided in [2]. It seems that in the present case, deterministic techniques are more appropriate; they are based on the decomposition of tensor arrays into elementary terms [3].

To be more explicit, a finite number of excitation and emission frequencies are measured, so that the measurements are stored in a finite array of order 3 and finite dimensions, say $I \times J \times K$:

$$T_{ijk} = I(\lambda_f(i), \lambda_e(j), k),$$

$1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K$. Tensor T can always be decomposed into a sum of elementary terms as:

$$T_{ijk} = \sum_{\ell=1}^{R} \lambda(\ell)\, A_{i\ell}\, B_{j\ell}\, C_{k\ell} \qquad (5)$$

where A, B and C are matrices with unit-norm columns, and where R is a sufficiently large integer. This can be referred to as a Polyadic decomposition of T [4]. The smallest integer R that can be found such that the equality above holds exactly is called the *tensor rank* [5]. For this value of R, the above decomposition is called the Polyadic Canonical decomposition (CP) of tensor T. It is clear, by comparing equations (4) and (5), that thanks to uniqueness of the CP, one can identify $\gamma_\ell(\lambda_f(i))$ with $A_{i\ell}$, $\epsilon_\ell(\lambda_e(j))$ with $B_{j\ell}$ and $c_{k,\ell}$ with $C_{k\ell}$. Hence, the computation of the CP yields emission spectra of each component as well as their concentration. There is no need to know in advance what are the components expected to be present in the solution.

This decomposition differs from the decomposition of matrices into a sum of rank-1 terms in several respects [1]. In particular, it is unique if the rank R of T is smaller than a known bound [3]. This is not the case for matrices, for which uniqueness can be achieved only thanks to orthogonality constraints imposed among the columns of A (resp. B), which leads to the Singular Value Decomposition (SVD). However, such a constraint has no physical meaning, and would not yield the spectra we are looking for.

Uniqueness of the CP is the main reason to resort to tensors rather than matrices. Note that in some scientific communities, the CP decomposition has received the name of "Parafac" [7] [6], which has no mathematical meaning. Such a terminology, introduced in the seventies by researchers in psychometrics, should be avoided, to the benefit of the more widely used acronym "CP" (even if often standing for "CanDecomp/Parafac" [3], to obtain agreement of all users).

Another nice property, which is not of crucial interest in the present framework, is that R is generally much larger than the smallest dimension of T. Such a property is very attractive in antenna array processing for instance [2], where linear mixtures may be "underdetermined". Tensor-based algorithms are then able to localize more radiating sources than sensors.

4. CONCLUSION

In this paper, we have outlined what are the goals we want to reach, what are the problems needing to be overcome, and what are the tools that we plan to use to solve them. Specific algorithms will be developed in order to cope with delays, with the positivity constraint of rank-1 tensors, and possibly with joint decomposition of mixtures of different nature. The cooperation initiated several years ago between I3S and PROTEE, is now more concrete thanks to the PhD of J.-P.Royer, launched in the frame of the PRES (Pole de Recherche et d'Enseignement Superieur) of the university of Nice. It gives the opportunity to I3S to participate more actively in the long-term study led by LRSAE [8] and PROTEE.

The expected impact is a better understanding of water exchanges (in particular underground) in the Var area, and a more efficient detection of polluting matters in water.

5. REFERENCES

[1] P. COMON, "Tensors, usefulness and unexpected properties", in *IEEE Workshop on Statistical Signal Processing* (SSP'09), Cardiff, UK, Aug. 31 - Sep. 3, 2009. invited keynote.

[2] P. COMON AND C. JUTTEN, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press, Oxford UK, Burlington USA, 2010. ISBN: 978-0-12-374726-6.

[3] P. COMON, X. LUCIANI, AND A. L. F. DE ALMEIDA, "Tensor decompositions, alternating least squares and other tales", *Jour. Chemometrics*, vol.23, 2009, pp. 393–405.

[4] F. L. HITCHCOCK, "The expression of a tensor or a polyadic as a sum of products", *J. Math. and Phys.*, vol.6, 1927, pp. 165–189.

[5] T. LICKTEIG, "Typical tensorial rank", *Linear Algebra Appl.*, vol.69, 1985, pp. 95–120.

[6] X. LUCIANI, S. MOUNIER, R. REDON, AND A. BOIS, "A simple correction method of inner filter effects affecting FEEM and its application to the Parafac decomposition", *Chemometrics and Intel. Lab. Syst.*, vol.96, 2009, pp. 227–238.

[7] A. SMILDE, R. BRO AND P. GELADI, *Multi-Way Analysis*, Wiley, 2004.

[8] G. FÉRAUD, C. POTOT, J.-F. FABRETTI, Y. GUGLIELMI, M. FIQUET, V. BARCI ET P.-CH. MARIA, "Trace elements as geochemical markers for surface waters and groundwaters of the Var catchment basin (Alpes Maritimes, France)", *C. R. Chimie*, vol.12, 2009, pp. 922–932.